International Journal of Information Management 34 (2014) 257–271

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Information management and improvement of citation indices

Valentin Gomez-Jauregui [a,*], Cecilia Gomez-Jauregui [b], Cristina Manchado [a,1], Cesar Otero [a,2]

[a] EGICAD Research Group, School of Civil Engineering, University of Cantabria, 39005 Santander, Spain
[b] Panda Security, Dpt. Panda Research, Gran Vía, 48001 Bilbao, Spain

ARTICLE INFO

Article history:
Available online 3 February 2014

Keywords:
Information management
Bibliometrics
Citation indices
Data cleaning
Software

ABSTRACT

Bibliometrics and citation analysis have become important sets of methods for library and information science, as well as exceptional sources of information and knowledge for many other areas. Their main sources are citation indices, which are bibliographic databases like Web of Science, Scopus, Google Scholar, etc. However, bibliographical databases lack perfection and standardization. There are several software tools that perform useful information management and bibliometric analysis importing data from them. A comparison has been carried out to identify which of them perform certain pre-processing tasks. Usually, they are not strong enough to detect all the duplications, mistakes, misspellings and variant names, leaving to the user the tedious and time-consuming task of correcting the data. Furthermore, some of them do not import datasets from different citation indices, but mainly from Web of Science (WoS).

A new software tool, called STICCI.eu (Software Tool for Improving and Converting Citation Indices – enhancing uniformity), which is freely available online, has been created to solve these problems. STICCI.eu is able to do conversions between bibliographical citation formats (WoS, Scopus, CSV, BibTex, RIS), correct the usual mistakes appearing in those databases, detect duplications, misspellings, etc., identify and transform the full or abbreviated titles of the journals, homogenize toponymical names of countries and relevant cities or regions and list the processed data in terms of the most cited authors, journals, references, etc.

1. Introduction

1.1. Bibliometrics and citation analysis

Bibliometrics and citation analysis have become very important sets of methods for library and information science in the last four decades, as well as exceptional sources of information and knowledge for many other areas. However, they are not new fields, according to Weinberg, which is cited in Hood and Wilson (2001). He claims that the first Hebrew citation indexes date from about the 12th century; Sengupta (also cited by Hood and Wilson) states that the first bibliometric study was produced by Campbell in 1896. Bibliometrics (English equivalent of the term 'bibliometrie', coined by Paul Otlet in 1934), was defined by Pritchard (1969) as "the application of mathematics and statistical methods to books and other media of communication"; It is a set of methods to quantitatively analyze scientific and technological literature. It permits the exploration of the impact of any research field, the influence of a group of researchers or institutes, the impact of a certain publication or the quantitative research of academic outputs. Other closed and related concepts are scientometrics (concerned with the quantitative features and characteristics of science) and informetrics, which are "a recent extension of the traditional bibliometric analyses also to cover non-scholarly communities in which information is produced, communicated, and used" (Ingwersen & Christensen, 1997), or more briefly, "quantitative methods in library, documentation and information science" (Egghe & Rousseau, 1990).

Citation analysis deals with the examination of the documents cited by scholarly works. Its main application was originally information retrieval and analyzing its quality. However, for the last years it has also been used for bibliometrics, evolving to evaluating and mapping researches, measuring the production and dissemination of scientific knowledge, becoming progressively more significant for assigning funding or career development, and also for establishing the journal impact factor. The main sources for citation analysis are citation indices, which are bibliographic databases that allows one to establish which later documents cite

* Corresponding author. Tel.: +34 942 206 757.
E-mail addresses: valen.gomez.jauregui@unican.es, valen.gomez.jauregui@gmail.com (V. Gomez-Jauregui), cecigj@gmail.com (C. Gomez-Jauregui), manchadoc@unican.es (C. Manchado), oteroc@unican.es (C. Otero).
[1] Tel.: +34 942 206 757.
[2] Tel.: +34 942 200 925.

0268-4012/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijinfomgt.2014.01.002

which earlier documents, which articles have been cited most frequently and who has cited them.

1.2. Citation indices and bibliographic databases

There are several citation indices, such as those published by Thomson Reuters' Institute for Scientific Information (ISI). Thomson Reuters' Web of Knowledge provides access to many bibliographic sources, such as MEDLINE and Web of Science (WoS), which collects many other citation indices. In total, they host a vast scholarly literature: 12,000 journals, 150,000 conference proceedings, 30,000 books and book chapters and 46 million records (Thomson Reuters, 2011).

Another important bibliographic database is SciVerse Scopus, published by Elsevier, which competes in completeness with WoS. Scopus claims to be the largest abstract and citation database of peer-reviewed research literature (Elsevier, 2011).

In comparison with these two citation indices by subscription, there are some other databases freely available: Google Scholar (with a vast information but with unreliable precision, as it will be discussed later), PubMed (specialized in life sciences and biomedical topics), CiteSeerX (the first automated citation indexing), Scirus (freely available search engine by Elsevier), getCITED (whose information is entered by members) or Microsoft Academic Search (academic search engine by Microsoft Research). Some other regional databases include SciELO, Dialnet, Latindex, etc. The fact of having multiple citation databases makes it necessary to compare them both from the scientometric and from the informetric points of view, by means of providing a set of measures for doing it systematically (Bar-Ilan, Levene, & Lin, 2007).

WoS and Scopus are the most reliable databases existing at the moment. In addition to their consistency, both citation indices offer a different selection of possibilities to export the records obtained by their respective search engines: plain text (.txt), tab-delimited (for Windows and Mac), comma separated values (CSV), web format (.html), BibTeX Bibliography Database (.bib), as well as for bibliographic management tools in Research Information Systems standardized format (.ris) like EndNote, Reference Manager, RefWorks, ProCite, etc. (Fig. 1).

Google Scholar (GS) and Microsoft Academic Search (MAS) cannot export their results directly, but only by means of Publish or Perish (Harzing, 2007), a very useful citation analysis software program, but with some limitations for certain tasks. For instance, although Publish or Perish can export a dataset obtained from GS or MAS to several output formats (BibTeX, CSV, EndNote, ISI and RefMan/RIS), it cannot export the full citations included in each work (Fig. 1c).

1.3. Comparison between WoS, Scopus and GS

A comparison between WoS and Scopus shows that "Scopus includes a more expanded spectrum of journals (...), and its citation analysis is faster and includes more articles than the citation analysis of WoS" (Falagas, Pitsouni, Malietzis, & Pappas, 2008). Scopus claims a worldwide coverage with more than 50% of its documents coming from Europe, Latin America and the Asian-Pacific Region (Elsevier, 2011) while WoS' contents are limited mainly to North America and Western European. Some other sources (Vieira & Gomes, 2009) confirm that Scopus provides the best coverage of social sciences literature, as well as for human–computer interaction literature, due to coverage of relevant ACM (Association for Computing Machinery) and IEEE (Institute of Electrical and Electronics Engineers) peer-reviewed conference proceedings. In addition, Scopus appears to have greater coverage of selected scientific areas, such as computer science, engineering, clinical medicine and biochemistry (Klavans & Boyack, 2007; Harzing, 2010). Finally, Scopus displays full cited reference information (although many times not in the same order), unlike WoS, which only displays first author, year, journal title, volume, first page number and doi. Compared to both of them, GS is the best one at coverage, it is increasing rapidly and it is more successful at retrieving citations (Harzing, 2010; Thornley, McLoughlin, Johnson, & Smeaton, 2011).

There are several authors (Bornmann, Leydesdorff, Walch-Solimena, & Ettl, 2011; Vieira & Gomes, 2009) confirming that Scopus has more errors, misspellings, inconsistencies and name variants (e.g. Munchen, München, Munich), at least compared to WoS. According to Lopez-Piñero (1992), cited in Postigo Jimenez, Díaz Casero, and Hernández Mogollón (2008), about 25% of the data obtained in the WoS were corrupted, while Postigo Jimenez et al. (2008) states that, for instance, during their bibliometric studies this number was sensibly reduce to approx. 18%. Vieira and Gomes (2009), analyzing inconsistencies in addresses, dates, volumes and issues (not in authors, for instance), found a total of 72 errors in 1965 documents of WoS (4%) and 258 errors in 1979 documents of Scopus (13%). What's more, errors in citations can reach up to 50% depending on the journal, with a minimum rate of about 10% (Libmann, 2007). Related to the other main database, Thornley et al. (2011) state that "GS could be an impractical tool for author searching" and that it lacks accuracy in its date fields, which could be a severe limitation. Harzing (2007) also confirms that "some references contain mixed-up fields (...) because its sources are inaccurate or difficult to parse automatically by Google's web crawler". The same author also confirms some other disadvantages of using GS: it includes some non-scholarly citations, not all scholarly journals are indexed in GS, its coverage might be uneven across different fields of study (e.g. the Natural and Health Sciences), it does not perform as well for older publications and GS automatic processing creates occasional nonsensical results (Harzing, 2008).

In conclusion, even though the number of mistakes is being reduced along the years, there are still many inconsistencies and errors that must be corrected.

1.4. Pre-processing datasets

Thus, we should stress the importance of pre-processing the data imported from any bibliographic database, in order to avoid mistakes, misspellings and inconsistencies. By doing so, the results of any citation metrics or bibliometric analysis would be more accurate, realistic and valid. Moreover, the difference between doing it manually or with specialized tools can be significant in terms of rapidity, efficiency and precision, which are the main problems to be addressed in this work.

A key aspect of pre-processing the data is to clean it and to purge it. This task could only be categorized as part of the de-duplication process, which for each entity in a database either merges the identified duplicate records into one combined record, or removes some records from the database until it only contains a single record for each entity. Duplicate records must be detected and corrected because, for instance, they could incorrectly increase the number of appearances of one item in a dataset, even if they are meant to be the same one. For example, if a search is made for finding the works made by a certain author (e.g. "Carro Pérez, Consuelo") in two different databases (e.g. in Scopus with 15 records found and in WoS with 17 records), and 13 of her works appear in both datasets, it would be desirable to delete or merge those 13 in the global combined dataset; the reason being that she would not be the author of 32 but of 19 publications and her h-index or g-index could be very different.

Another example, according to Thornley et al. (2011) would be the case of a certain work that could be found in several versions: